DFTT 11/00
arXiv:hep-ex/0003001

What is the Usefulness of Frequentist Confidence Intervals?

Carlo Giunti
INFN, Sez. di Torino, and Dip. di Fisica Teorica, Univ. di Torino, I–10125 Torino, Italy
(March 1, 2000)

Abstract

The following questions are discussed: "Why confidence intervals are a hot topic?"; "Are confidence intervals objective?"; "What is the usefulness of coverage?"; "How to obtain useful information from experiment?"; "The confidence level must be chosen independently from the knowledge of the data?".

PACS numbers: 06.20.Dk

The problem of getting meaningful information from the statistical analysis of experimental data in high energy physics has attracted recently great attention, reaching a (local) maximum at the Workshop on "Confidence Limits" held at CERN in January [1]. Having participated to that Workshop and having read some of the related material available at the Workshop Web page [1], I think that there is a certain confusion on the usefulness of Frequentist confidence intervals and some clarifications are necessary.

In the following I will consider and answer some crucial questions in the framework of the Frequentist theory of statistical inference. I will assume that the reader is familiar with the theory and its problems (if not, see [1–20]).

1. **Why confidence intervals are a hot topic?**

   The current debate on the methods of statistical analysis of experimental data follows mainly from the proposal at the end of 1997 of the the Unified Approach by Feldman and Cousins [6] and its immediate adoption as the recommended Frequentist method in the 1998 edition of the *Review of Particle Physics* (RPP) of the Particle Data Group (PDG) [7]. Although it may be true that "there is no PDG method" [21], it is a matter of fact that RPP is a guide for the Physics community. Most physicists have faith in what is written in RPP, especially regarding the fields in which they are not experts (unfortunately most human beings can achieve expertise in one or a few fields and it is naturally correct to believe to experts in the other fields if nothing that they say is obviously wrong). Therefore, the authors of RPP have a responsibility for what they write.

   The immediate adoption of the Unified Approach by the PDG has been considered by many rather premature, taking into account that it happened before testing the performances of the Unified Approach in real experiments. These concerns received dramatic confirmations from the unphysical results obtained in two of the first applications [22,23]. Several papers [8,11–14] have followed the one by Feldman and Cousins, proposing alternative Frequentist methods. Hence, at present there are several Frequentist approaches available and each analyzer of experimental data must choose one among them independently of the knowledge of the data, in order to preserve the property of coverage [6]. One of the main issues in the present debate on confidence intervals is the study of the properties of the different Frequentist methods in order to allow a meaningful choice of the method to be used in a practical application.

   Another problem with the 1998 edition of RPP is that the emphasis on the Unified Approach is likely correlated with the disappearance of the useful description of the Bayesian approach present in the 1996 edition of RPP [5]. It seems hard to argue that this is not a biased choice.

2. **Are confidence intervals objective?**

   It is well known that credibility intervals obtained in the framework of the Bayesian theory (see, for example, [24]) are subjective because of the necessity to have a prior probability distribution function for the quantity under measurement.

   In the Frequentist–Bayesian debate some experts biased towards the Frequentist approach say that they want to know what was measured in the experiment, without the

subjective Bayesian prior of the experimenter. But also in the Frequentist approach the experimenter must choose a method to construct the confidence belt and the result that he will obtain depends on this choice (in Ref. [16] it has been shown that Frequentist confidence intervals are objective only from a statistical point of view). Thus, it is clear that Frequentist confidence intervals (as Bayesian credibility intervals) do not describe what is measured independently from subjective choices!

Actually, in the framework of the Frequentist theory there is no way to get a result without the subjective choice of the method to construct the confidence belt (approximate methods as maximum likelihood also need subjective choices). On the other hand, working in the framework of the Bayesian theory, one can present the likelihood function as the result of the experiment (or its normalized version called "relative belief updating ratio" [25]) and everyone can obtain a credibility interval using her/his prior. Moreover, the Bayesian prior takes into account in a proper way the subjective believe that may come from a solid experience in the field, whereas the choice of a specific Frequentist method seems much more arbitrary.

Since the Frequentist confidence intervals are subjective as the Bayesian credibility intervals, but Bayesian theory takes into account subjective belief based on experience in a proper way, I think that, *contrary to what is usually believed, a choice of the method based on subjectivity favors the Bayesian approach*. A choice of the Frequentist approach is reasonable only if based on the main property of Frequentist confidence intervals, *coverage*, that I will discuss in the following items.

3. **What is the usefulness of coverage?**

   Some experts say that coverage implies that in order to be right, for example, 90% of the times, 10% of the times you must get a wrong result and you must give it, even if you know (or have a strong suspicion) that it is wrong (for example, an empty confidence interval). If I tell this to any pedestrian, he will think that I am nuts: if I know that the result is wrong why should I give it? It is not only useless, it may also confuse other people. So, I think that among reasonable people we can agree that *wrong results are useless and should not be given*.

   I think that coverage is useful because one knows the probability, given by the confidence level, that the confidence interval covers the true value of the quantity under measurement. Each confidence interval obtained in an experiment has this property, independently from the results and even existence of other experiments. Therefore, there is no need to give wrong confidence intervals!

   In principle, if one could make many experiments to measure a physical quantity μ, each experiment producing a confidence interval with a chosen confidence level, one could collect all the confidence intervals (including those that are known to be wrong), producing a set of intervals that cover the true value of μ with a probability given by the confidence level. But in practice, at least in high energy physics, there are only a few experiments (sometimes one or two) that measure each physical quantity. Therefore, the set of confidence intervals is too small to be of any usefulness. Instead, one is interested to get useful information from each experiment.

4. **How to obtain useful information from experiment?**

I think that a procedure that allows to get always useful information from an experiments is the following:

(a) Choose the Frequentist method with the desired properties independently of the knowledge of the data (see [16]).

(b) If the data do not indicate any unlikely statistical fluctuation and the confidence interval obtained with the chosen Frequentist method looks fine, the confidence interval can be given and one knows that it covers the true value of the quantity under measurement with a probability given by the confidence level.

(c) If it is clear that the data indicate an unlikely statistical fluctuation (as less events than the expected background measured in a Poisson process with known background) and the confidence interval obtained with the chosen Frequentist method is suspected to be wrong (for example, too small or even vanishing), the confidence interval should not be given.

Feldman and Cousins [6] proposed that in such cases the experimenter should give also what they called "sensitivity", but should be called more appropriately *"exclusion potential"* [16], because it is calculated assuming that the quantity under measurement is zero. However, since the exclusion potential cannot be combined with the confidence interval that has been obtained in the experiment, it is not clear what is the usefulness of giving two quantities instead of one, except as a warning that the confidence interval is likely to be wrong. But in this case it is better not to give it! Two quantities produce only confusion if one of them tells that the other is not reliable. Thus, the solution proposed by Feldman and Cousins (recommended also in the 1998 edition of RPP [7]) to give two quantities instead of one is just the opposite of what it is reasonable to do: give nothing! (Here I am discussing only Frequentist quantities. One can always give Bayesian quantities, as discussed in the next item.)

(d) In any case the experimenters should analyze their data using the Bayesian theory, that allows always to obtain meaningful results. The experimenters can give the likelihood function or the relative belief updating ratio [25], that represent the objective result of the experiment, and can give also a credibility interval obtained with their prior based on experience and knowledge of the experiment.

Following this procedure, experiments will always produce a result in the framework of the Bayesian theory and will produce also a Frequentist result only if it is a reliable one.

As an illustration, let us consider the well known case of the KARMEN experiment on the search for short-baseline $\bar{\nu}_\mu \to \bar{\nu}_e$ oscillations [22]. In the middle of 1998 the KARMEN collaboration reported the observation of zero events in a Poisson process with a known background of 2.88 ± 0.13 events [22]. Using the Unified Approach they obtained an exclusion curve in the space of the neutrino mixing parameters that seemed to exclude almost all the region allowed by the positive results of the LSND

$\bar{\nu}_\mu \to \bar{\nu}_e$ experiment [26]. The exclusion curve of the KARMEN experiment lead many people to believe that the LSND evidence in favor of $\bar{\nu}_\mu \to \bar{\nu}_e$ oscillations was almost ruled out by the result of the KARMEN experiment, in spite of the fact that the exclusion potential of the KARMEN experiment was about four times larger than the actual upper limit. This discrepancy was due to the observation of less events than the expected background. In 1999 the KARMEN experiment reported the observation of as many events as expected from background [27], resulting in an upper limit practically coincident with the exclusion potential, compatible with the results of the LSND experiment (see also [28]). It is clear that the result presented in 1998 has been worse than useless: its only effect has been to confuse people. This confusion could have been avoided if the KARMEN collaboration would not have presented the 1998 exclusion curve obtained with the Unified Approach, since it was clearly meaningless from a physical point of view (although statistically correct if the KARMEN collaboration did not choose the Unified Approach on the basis of the data, for example because it gave a bound more stringent than other methods). Moreover, the KARMEN collaboration did not [22] (and still does not, whereas curiously they continue to give the useless 1998 exclusion curve [27]) present the result of a Bayesian analysis, which is less sensitive to fluctuations of the background than the Unified Approach (see, for example, [29,16]). This is probably a negative consequence of the above mentioned (question 1) bias of the 1998 edition of RPP towards the Unified Approach.

Another lesson to be learned from the KARMEN example regards the usefulness of the goodness of fit test proposed by Feldman and Cousins [6]: in the Poisson with background case the natural analogue for the goodness of fit is the probability to obtain $n \leq n_{\rm obs}$ under the best-fit assumption of $\mu = 0$, where μ is the mean of signal events, n is the number of events and $n_{\rm obs}$ is the number of observed events. In the case of the KARMEN experiment $n_{\rm obs}$ was zero in 1998 and the probability to obtain $n = 0$ with $\mu = 0$ and a known background of 2.88 ± 0.13 events is 5.6%. This probability is not unacceptable, but imagine that we decide to reject a goodness of fit lower than 10%. The problem is that in this case there is nothing to reject, because the background is known [30] and low fluctuations of the background are allowed. Thus, in the case of a Poisson process with know background the goodness of fit test proposed by Feldman and Cousins is useless.

5. **The confidence level must be chosen independently from the knowledge of the data?**

   As far as I know, this important question has not been discussed in the literature. I think that the answer depends on the use that one is going to make with the confidence intervals.

   If many experiments have been done, the resulting confidence intervals at a certain confidence level form a set that covers the true value of the quantity under measurement with probability given by the confidence level. This property is damaged if the confidence level is chosen on the basis of the data. For example, in the case of a Poisson process with known background, it is reasonable to choose a higher confidence level if less events than the expected background have been observed, because the low fluctua-

tion of the background induces a certain skepticism on the reliability of the confidence interval. But in this way the set of confidence intervals with high confidence level is unbalanced towards low values of the quantity under measurement, whereas the set of confidence intervals with low confidence level is unbalanced towards high values of the quantity under measurement. Thus, the sets of confidence intervals do not have correct coverage and the answer to the question above is "yes".

In practice, however, at least in high energy physics research, one does not have the possibility to do many experiments for the measurement of a certain quantity and one is not interested in collecting a set of confidence intervals that cover the true value of the quantity under measurement with a given probability. As discussed above in 3, each experimental collaboration is interested to obtain a meaningful and reliable result in its experiment. Nobody is going to collect sets of confidence intervals obtained in different experiments and study their properties. The confidence interval obtained in each experiment is considered individually, not embedded in a set. In this case the confidence level can be chosen after the data are known without spoiling coverage.

It is now well known that the method to construct the confidence belt must be chosen independently of the knowledge of the data [6]. A simple reason is that knowing the data one can always construct a confidence belt that gives any wanted confidence interval at some (sometimes small) confidence level. But when the method to construct the confidence belt has been chosen the coverage of the confidence belt is guaranteed for any value of the confidence level and the freedom to choose the confidence level does not allow one to get a wanted confidence interval.

Taking into account that the confidence interval can be chosen at will, I think that it is highly desirable that the experimental collaborations publish not a single confidence interval at an arbitrary confidence level, but the entire *confidence distribution* of the parameter [31], *i.e.* the limits of the confidence interval as functions of the confidence level, at least for large values of the confidence level (say larger than 68%). In these days this can be easily done even for multi-dimensional confidence intervals by giving a table available as a file through the Internet and/or an interpolating function.

In conclusion, I have discussed some crucial questions regarding Frequentist confidence intervals. I hope that the answers that I have given will at least stimulate the debate on the subject and, if they are right, will contribute to an improvement of the understanding of the usefulness of confidence intervals.

REFERENCES

[1] Workshop on "Confidence Limits", CERN, 17-18 Jan. 2000, http://www.cern.ch/-CERN/Divisions/EP/Events/CLW.

[2] J. Neyman, Philos. Trans. R. Soc. London Sect. A **236**, 333 (1937), reprinted in *A selection of Early Statistical Papers on J. Neyman*, University of California, Berkeley, 1967, p. 250.

[3] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, *Statistical Methods in Experimental Physics*, North Holland, Amsterdam, 1971.

[4] R.D. Cousins, Am. J. Phys. **63**, 398 (1995).

[5] R.M. Barnett *et al.* (Particle Data Group), Phys. Rev. D **54**, 1 (1996).

[6] G.J. Feldman and R.D. Cousins, Phys. Rev. D **57**, 3873 (1998), arXiv:physics/9711021.

[7] C. Caso *et al.* (Particle Data Group), Eur. Phys. J. C **3**, 1 (1998).

[8] C. Giunti, Phys. Rev. D **59**, 053001 (1999), arXiv:hep-ph/9808240.

[9] C. Giunti, Phys. Rev. D **59**, 113009 (1999), arXiv:hep-ex/9901015.

[10] C. Giunti, in *Summary of the NOW'98 Phenomenology Working Group*, http://www.-nikhef.nl/pub/conferences/now98/presentations.html [arXiv:hep-ph/9906251].

[11] S. Ciampolillo, Il Nuovo Cimento A **111**, 1415 (1998).

[12] B.P. Roe and M.B. Woodroofe, Phys. Rev. D **60**, 053009 (1999), arXiv:physics/9812036.

[13] M. Mandelkern and J. Schultz, arXiv:hep-ex/9910041.

[14] G. Punzi, arXiv:hep-ex/9912048.

[15] R.D. Cousins, arXiv:physics/0001031.

[16] C. Giunti and M. Laveder, arXiv:hep-ex/0002020.

[17] R.D. Cousins, talk presented at the CERN Workshop on "Confidence Limits" [1], http://www.cern.ch/CERN/Divisions/EP/Events/CLW/PAPERS/PS/zech.ps.

[18] C. Giunti, arXiv:hep-ex/0002042, talk presented at the CERN Workshop on "Confidence Limits" [1].

[19] G. Zech, talk presented at the CERN Workshop on "Confidence Limits" [1], http://-www.cern.ch/CERN/Divisions/EP/Events/CLW/PAPERS/PS/cousins.ps.

[20] "Panel Discussion" at the CERN Workshop on "Confidence Limits" [1], http://www.-cern.ch/CERN/Divisions/EP/Events/CLW/QA/PS/clwdiscuss.ps.

[21] Don Groom in [20].

[22] K. Eitel and B. Zeitnitz (KARMEN coll.), Nucl. Phys. B (Proc. Suppl.) 77, 212 (1999), arXiv:hep-ex/9809007 [32].

[23] L. Baudis *et al.* (Heidelberg-Moscow coll.), Phys. Rev. Lett. **83**, 41 (1999), arXiv:hep-ex/9902014.

[24] G. D'Agostini, CERN Yellow Report 99-03 (available at http://www-zeus.roma1.infn.-it/%7Eagostini/prob+stat.html).

[25] P. Astone and G. D'Agostini, arXiv:hep-ex/9909047; G. D'Agostini, Am. J. Phys. **67**, 1260 (1999) [arXiv:physics/9908014]; arXiv:physics/9906048; G. D'Agostini and G. Degrassi, Eur. Phys. J. C **10**, 663 (1999).

[26] C. Athanassopoulos *et al.* (LSND coll.), Phys. Rev. Lett. **75**, 2650 (1995); Phys. Rev. Lett. **77**, 3082 (1996); Phys. Rev. Lett. **81**, 1774 (1998); G. Mills (LSND coll.), Talk presented at the XXXIV$^{\text{th}}$ Rencontres de Moriond *Electroweak Interactions and Unified Theories*, Les Arcs, March 1999 (http://moriond.in2p3.fr/EW/transparencies).

[27] T.E. Jannakos (KARMEN coll.), arXiv:hep-ex/9908043 [32].
[28] K. Eitel, New Journal of Physics **2**, 1.1 (2000), arXiv:hep-ex/9909036 [32].
[29] P. Astone and G. Pizzella, arXiv:hep-ex/0002028, talk presented at the CERN Workshop on "Confidence Limits" [1].
[30] B. Armbruster (KARMEN coll.), Contribution to the Neutrino Oscillation Workshop NOW'98, 7-9 September 1998, Amsterdam, http://www.nikhef.nl/pub/conferences/-now98/papers.html [32].
[31] D.R. Cox, Ann. Math. Statist. **29**, 357 (1958).
[32] KARMEN WWW page: http://www-ik1.fzk.de/www/karmen.